

Optimal spectral norm rates for noisy low-rank matrix completion

Karim Lounici*

*School of Mathematics
Georgia Institute of Technology
Atlanta, GA 30332-0160
e-mail: klounici@math.gatech.edu*

Abstract: In this paper we consider the trace regression model where n entries or linear combinations of entries of an unknown $m_1 \times m_2$ matrix A_0 corrupted by noise are observed. We establish for the nuclear-norm penalized estimator of A_0 introduced in [13] a general sharp oracle inequality with the spectral norm for arbitrary values of n, m_1, m_2 under an incoherence condition on the sampling distribution Π of the observed entries. Then, we apply this method to the matrix completion problem. In this case, we prove that it satisfies an optimal oracle inequality for the spectral norm, thus improving upon the only existing result [13] concerning the spectral norm, which assumes that the sampling distribution is uniform. Note that our result is valid, in particular, in the high-dimensional setting $m_1 m_2 \gg n$. Finally we show that the obtained rate is optimal up to logarithmic factors in a minimax sense.

AMS 2000 subject classifications: Primary 62J99, 62H12; secondary 60B20, 60G15.

Keywords and phrases: matrix completion, low-rank matrix estimation, spectral norm, optimal rate of convergence, noncommutative Bernstein inequality, Lasso.

1. Introduction

Consider n independent observations $(X_i, Y_i), i = 1, \dots, n$, satisfying the trace regression model:

$$Y_i = \text{tr}(X_i^\top A_0) + \xi_i, \quad i = 1, \dots, n, \quad (1.1)$$

where X_i are random matrices with dimensions $m_1 \times m_2$, Y_i are random variables in \mathbb{R} , $A_0 \in \mathbb{R}^{m_1 \times m_2}$ is an unknown matrix, $\xi, \xi_i, i = 1, \dots, n$ are i.i.d. zero mean random variables with $\sigma_\xi^2 = \mathbb{E}\xi^2 < \infty$ and $\text{tr}(B)$ denotes the trace of matrix B . We consider the problem of estimation of A_0 based on the observations $(X_i, Y_i), i = 1, \dots, n$.

For any matrices $A, B \in \mathbb{R}^{m_1 \times m_2}$, we define the scalar products

$$\langle A, B \rangle = \text{tr}(A^\top B),$$

*supported in part by NSF grant DMS 1106644 and Simons foundation grant 209842.

and

$$\langle A, B \rangle_{L_2(\Pi)} = \frac{1}{n} \sum_{i=1}^n \mathbb{E}(\langle A, X_i \rangle \langle B, X_i \rangle).$$

Here $\Pi = \frac{1}{n} \sum_{i=1}^n \Pi_i$, where Π_i denotes the distribution of X_i . The corresponding norm $\|A\|_{L_2(\Pi)}$ is given by

$$\|A\|_{L_2(\Pi)}^2 = \frac{1}{n} \sum_{i=1}^n \mathbb{E}(\langle A, X_i \rangle^2).$$

Example 1: Matrix completion. Let the design matrices X_i be i.i.d. with distribution Π on the set

$$\mathcal{X} = \{e_j(m_1)e_k^\top(m_2), 1 \leq j \leq m_1, 1 \leq k \leq m_2\}, \quad (1.2)$$

where $e_k(m)$ are the canonical basis vectors in \mathbb{R}^m . The set \mathcal{X} forms an orthonormal basis in the space of $m_1 \times m_2$ matrices that will be called the matrix completion basis. Let also $n < m_1 m_2$. Then the problem of estimation of A_0 coincides with the problem of matrix completion with random sampling distribution Π . Existing results typically assume that Π is the uniform distribution on \mathcal{X} . See, for instance, [9, 19] for the non-noisy case ($\xi_i = 0, i = 1, \dots, n$) and [13] for the noisy case and the references cited therein. In several applications, like the Netflix problem, the distribution Π is not necessarily uniform on \mathcal{X} . We will show that optimal estimation of A_0 is possible in this context under a weaker set of conditions as compared to those used in [7, 9, 19]. One can also consider other matrix measurement models. For instance, [10] considers sampling without replacement in the set \mathcal{X} defined in (1.2) and [11] investigates several orthonormal families in the context of Quantum tomography.

Example 2. Column masks. Let the design matrices X_i be independent matrices, which have only one nonzero column. The trace regression model can be then reformulated as a longitudinal regression model, with different distributions of X_i corresponding to different tasks; see [1, 15, 21] for more details and the references cited therein.

Example 3. "Complete" subgaussian design. Let the design matrices X_i are i.i.d. replications of a random matrix X such that $\langle A, X \rangle$ is a subgaussian random variable for any $A \in \mathbb{R}^{m_1 \times m_2}$. This approach originates from compressed sensing, where typically the entries of X are either i.i.d. standard Gaussian or Rademacher random variables. The problem of exact reconstruction of A_0 under such a design in the non-noisy setting was studied in [5, 16, 20], whereas estimation of A_0 in the presence of noise is analyzed in [5, 16, 21], among which [5, 11, 21] treat the high-dimensional case $m_1 m_2 > n$.

We consider the following procedure introduced recently in [13]

$$\hat{A}^\lambda = \operatorname{argmin}_{A \in \mathbb{R}^{m_1 \times m_2}} \left\{ \|A\|_{L_2(\Pi)}^2 - \left\langle \frac{2}{n} \sum_{i=1}^n Y_i X_i, A \right\rangle + \lambda \|A\|_1 \right\}, \quad (1.3)$$

where $\lambda > 0$ is the regularization parameter and $\|A\|_1$ is the nuclear norm of A .

In the matrix completion problem, the sampling scheme is typically assumed to be uniform on \mathcal{X} and this assumption is crucial to establish the exact recovery in the noiseless case or to derive the optimal rates of estimation with the Frobenius norm in the setting $n < m_1 m_2$; see for instance [6, 9, 13] and the references cited therein. However, in several applications such as the Netflix problem, the practitioner does not choose the sampling scheme and the observed entries of A_0 are not guaranteed to follow the uniform distribution. Therefore, the existing exact recovery or estimation results do not cover this situation.

In this paper, we concentrate mainly on the matrix completion problem. Our contributions are the following. First, we establish for the estimator (1.3) the following result. If A_0 is low rank, Π satisfies an incoherence condition and, in addition, some additional mild conditions are satisfied, then we have for any $t > 0$ with probability at least $1 - e^{-t}$

$$\|\hat{A}^\lambda - A_0\|_\infty \leq C(\sigma \vee a) \sqrt{m_1 m_2} \sqrt{(m_1 \vee m_2) \frac{t + \log(m_1 + m_2)}{n}}, \quad (1.4)$$

where $C > 0$ is a numerical constant, a is a bound on the absolute values of the entries of A_0 and $\|\cdot\|_\infty$ is the spectral norm. Second, we show that the above rate is optimal (in the minimax sense) up to logarithmic factors on a particular class of low rank matrices.

Note that the existing estimation results concern usually the Frobenius norm [5, 10, 11, 18]. The only existing estimation result for the spectral norm is due to [13] which assumes that the entries are sampled uniformly at random. In this case, the estimator (1.3) can be computed directly by soft-thresholding of the singular values in the SVD of $\mathbb{X} = \frac{m_1 m_2}{n} \sum_{i=1}^n Y_i X_i$ (see Equation (3.2) in [13]). Exploiting this explicit simple form, [13] established (1.4) for the procedure (1.3). This approach does not generalize to other sampling distribution Π since (1.3) does not admit an explicit form in general. In this paper, we propose an alternative approach to derive for the estimator (1.3) the oracle inequality (1.4) when the sampling distribution Π satisfies an incoherence condition, which covers in particular the case of uniform sampling Π and also holds in more general situations.

Note finally that the results of this paper are obtained for general settings of n, m_1, m_2 . In particular they are valid in the high-dimensional setting, which corresponds to $m_1 m_2 \gg n$, with low rank matrices A_0 .

In section 2, we recall some tools and definitions and establish a preliminary result. In Section 3, we establish a general oracle inequality for the spectral norm. In Section 4, we apply the general result of the previous section to the matrix completion problem and establish the optimality (up to logarithmic factors) of (1.3). Finally, Section 5 contains additional material and proofs.

2. Tools and preliminary result

We recall first some basic facts about matrices. Let $A \in \mathbb{R}^{m_1 \times m_2}$ be a rectangular matrix, and let $r = \text{rank}(A) \leq \min(m_1, m_2)$ denote its rank. The singular value decomposition (SVD) of A admits the form

$$A = \sum_{j=1}^r \sigma_j(A) u_j^{(A)} \otimes v_j^{(A)},$$

with orthonormal vectors $u_1^{(A)}, \dots, u_r^{(A)} \in \mathbb{R}^{m_1}$, orthonormal vectors $v_1^{(A)}, \dots, v_r^{(A)} \in \mathbb{R}^{m_2}$ and real numbers $\sigma_1(A) \geq \dots \geq \sigma_r(A) > 0$ (the singular values of A). The pair of linear vector spaces $(S_1(A), S_2(A))$ where $S_1(A)$ is the linear span of $\{u_1^{(A)}, \dots, u_r^{(A)}\}$ and $S_2(A)$ is the linear span of $\{v_1^{(A)}, \dots, v_r^{(A)}\}$ will be called the *support* of A . We will denote by $S_j(A)^\perp$ the orthogonal complements of $S_j(A)$, $j = 1, 2$, and by P_S the orthogonal projector onto the linear vector subspace S of \mathbb{R}^{m_j} , $j = 1, 2$. For any $A \in \mathbb{A}$ with support (S_1, S_2) , we define

$$\mathcal{P}_A(B) := B - P_{S_1^\perp} B P_{S_2^\perp}, \quad \mathcal{P}_A^\perp(B) := P_{S_1^\perp} B P_{S_2^\perp}, \quad B \in \mathbb{R}^{m_1 \times m_2}.$$

The Schatten- p (quasi-)norm $\|A\|_p$ of matrix A is defined by

$$\|A\|_p = \left(\sum_{j=1}^{\min(m_1, m_2)} \sigma_j(A)^p \right)^{1/p} \quad \text{for } 0 < p < \infty, \quad \text{and} \quad \|A\|_\infty = \sigma_1(A).$$

Recall the well-known *trace duality* property:

$$|\text{tr}(A^\top B)| \leq \|A\|_1 \|B\|_\infty, \quad \forall A, B \in \mathbb{R}^{m_1 \times m_2}.$$

We will also use the fact that the subdifferential of the convex function $A \mapsto \|A\|_1$ is the following set of matrices:

$$\partial \|A\|_1 = \left\{ \sum_{j=1}^r u_j^{(A)} \otimes v_j^{(A)} + P_{S_1(A)^\perp} W P_{S_2(A)^\perp} : \|W\|_\infty \leq 1 \right\} \quad (2.1)$$

(cf. [24]).

We will need the following quantities introduced in [12]

$$\kappa_r = \kappa_r(\Pi) := \inf \{ \|B_2\|_{L_2(\Pi)} : B \in \mathbb{R}^{m_1 \times m_2}, \|B\|_2 = 1, \text{rank}(B) \leq r \}$$

and

$$\kappa'_r = \kappa'_r(\Pi) := \sup \{ \|B_2\|_{L_2(\Pi)} : B \in \mathbb{R}^{m_1 \times m_2}, \|B\|_2 = 1, \text{rank}(B) \leq r \}.$$

These quantities $\kappa_r(\Pi)$ and $\kappa'_r(\Pi)$ measure the "distorsion" on the set of low rank matrices between the geometries induced respectively by the $L_2(\Pi)$ and Frobenius norms.

We introduce the following measure of coherence

$$\rho = \rho(\Pi) := \sup \left\{ \frac{|\langle A, B \rangle_{L_2(\Pi)}|}{\|A\|_1 \|B\|_1} : \forall A, B \in \mathbb{R}^{m_1 \times m_2}, \langle A, B \rangle = 0 \right\}. \quad (2.2)$$

We can now state our incoherence condition.

Assumption 1. Let $c_0 \geq 0$, $\alpha > 1$ and $r \geq 1$. We have

$$\rho \leq \frac{\kappa_1^2}{(1 + 2c_0)\alpha r},$$

The quantity ρ is the natural extension to the matrix case of the incoherence measure introduced for the sparse vector case in [8] and further studied in [2, 3, 14] and the references cited therein. Concerning the matrix completion problem, [5, 6, 9, 13] study the case of uniform at random sampling. Assumption 1 is then trivially satisfied with $\rho = 0$, since we have $\langle A, B \rangle_{L_2(\Pi)} = \frac{1}{m_1 m_2} \langle A, B \rangle$ for any $A, B \in \mathbb{R}^{m_1 \times m_2}$. Note also that [5, 6, 9] need in addition the following condition in order to recover A_0 in the noiseless case

$$\max_{u \in \mathcal{X}} |\mathcal{P}_{S_1, S_2}(u)|_2^2 \leq \frac{2\nu r}{m_1 \wedge m_2}, \quad \left| \sum_{j=1}^r u_j^{(A_0)} \otimes v_j^{(A_0)} \right|_\infty \leq \frac{2\nu r}{(m_1 \wedge m_2)^2},$$

for some $\nu > 0$ where $S_j = S_j(A_0)$, $j = 1, 2$ and $|\cdot|_2, |\cdot|_\infty$ denote respectively the l_2 and l_∞ vector norms. Although called "incoherence condition" in [9], this condition is entirely different from Assumption 1 and we do not need it to establish our estimation result.

In [13], the authors establish an oracle inequality for the $L_2(\Pi)$ norm under a condition akin to the restricted eigenvalue condition in sparse vector estimation: $\mu_{c_0}(A_0) < \infty$ for some $c_0 \geq 0$ where

$$\mu_{c_0}(A_0) := \inf \left\{ \mu > 0 : \|\mathcal{P}_{A_0}(B)\|_2 \leq \mu \|B\|_{L_2(\Pi)}, \forall B \in \mathbb{C}_{A_0, c_0} \right\},$$

and \mathbb{C}_{A_0, c_0} is the following cone of matrices

$$\mathbb{C}_{A_0, c_0} := \left\{ B \in \mathbb{R}^{m_1 \times m_2} : \|\mathcal{P}_{A_0}^\perp(B)\|_1 \leq c_0 \|\mathcal{P}_{A_0}(B)\|_1 \right\}.$$

Note that $\mu_{c_0}(A_0)$ is a nondecreasing function of c_0 . We establish in Proposition 1 below that Assumption 1 implies $\mu_{c_0}(A_0) < \frac{1}{\kappa_1} \sqrt{\frac{\alpha}{\alpha-1}}$ if $\text{rank}(A_0) \leq r$.

Proposition 1. Let Assumption 1 be satisfied for some $c_0 \geq 0$, $\alpha > 1$ and $r \geq 1$. Assume furthermore that $\kappa_1 = \kappa_1(\Pi) > 0$. Then, for any $A \in \mathbb{R}^{m_1 \times m_2}$ with $\text{rank}(A) \leq r$, we have

$$\mu_{c_0}(A) \leq \frac{1}{\kappa_1} \sqrt{\frac{\alpha}{\alpha-1}} < \infty.$$

Proof. We have

$$\begin{aligned}
 \|\mathcal{P}_A(B) + \mathcal{P}_A^\perp(B)\|_{L_2(\Pi)}^2 &= \|\mathcal{P}_A(B)\|_{L_2(\Pi)}^2 + \|\mathcal{P}_A^\perp(B)\|_{L_2(\Pi)}^2 + 2\langle \mathcal{P}_A(B), \mathcal{P}_A^\perp(B) \rangle_{L_2(\Pi)} \\
 &\geq \|\mathcal{P}_A(B)\|_{L_2(\Pi)}^2 + \|\mathcal{P}_A^\perp(B)\|_{L_2(\Pi)}^2 - 2\rho \|\mathcal{P}_A(B)\|_1 \|\mathcal{P}_A^\perp(B)\|_1 \\
 &\geq \|\mathcal{P}_A(B)\|_{L_2(\Pi)}^2 - 2\rho c_0 \|\mathcal{P}_A(B)\|_1^2 \\
 &\geq \|\mathcal{P}_A(B)\|_{L_2(\Pi)}^2 - 2\rho c_0 r \|\mathcal{P}_A(B)\|_2^2.
 \end{aligned} \tag{2.3}$$

Next, we treat $\|\mathcal{P}_A(B)\|_{L_2(\Pi)}^2$. For the sake of brevity, we set $r = \text{rank}(\mathcal{P}_A(B))$ and, for any $1 \leq j \leq r$, $\sigma_j = \sigma_j(\mathcal{P}_A(B))$, $u_j = u_j^{(\mathcal{P}_A(B))}$ and $v_j = v_j^{(\mathcal{P}_A(B))}$. Recall that the SVD of $\mathcal{P}_A(B)$ is

$$\mathcal{P}_A(B) = \sum_{j=1}^r \sigma_j u_j \otimes v_j.$$

For any $B \in \mathbb{R}^{m_1 \times m_2}$, we have

$$\begin{aligned}
 \|\mathcal{P}_A(B)\|_{L_2(\Pi)}^2 &= \left\| \sum_{j=1}^r \sigma_j u_j \otimes v_j \right\|_{L_2(\Pi)}^2 \\
 &= \sum_{j=1}^r \sigma_j^2 \|u_j \otimes v_j\|_{L_2(\Pi)}^2 + \sum_{j,k=1:j \neq k}^r \sigma_j \sigma_k \langle u_j \otimes v_j, u_k \otimes v_k \rangle_{L_2(\Pi)} \\
 &\geq \sum_{j=1}^r \sigma_j^2 \|u_j \otimes v_j\|_{L_2(\Pi)}^2 - \rho \sum_{j,k=1:j \neq k}^r \sigma_j \sigma_k \\
 &\geq \sum_{j=1}^r \sigma_j^2 \|u_j \otimes v_j\|_{L_2(\Pi)}^2 - \rho \left(\sum_{j=1}^r \sigma_j \right)^2 \\
 &\geq (\kappa_1^2 - \rho r) \sum_{j=1}^r \sigma_j^2 = (\kappa_1^2 - \rho r) \|\mathcal{P}_A(B)\|_2^2.
 \end{aligned} \tag{2.4}$$

Combining (2.3) and 2.4 with Assumption 1 yields

$$\begin{aligned}
 \|B\|_{L_2(\Pi)}^2 &\geq (\kappa_1^2 - \rho(1 + 2c_0)r) \|\mathcal{P}_A(B)\|_2^2 \\
 &\geq \frac{\kappa_1^2(\alpha - 1)}{\alpha} \|\mathcal{P}_A(B)\|_2^2.
 \end{aligned}$$

Thus, we get the result. \square

3. General oracle inequalities for the spectral norm

Define the random matrices

$$\mathbf{M}_1 = \frac{1}{n} \sum_{i=1}^n \xi_i X_i, \quad \mathbf{M}_2 = \frac{1}{n} \sum_{i=1}^n \langle A_0, X_i \rangle - \mathbb{E}(\langle A_0, X_i \rangle). \tag{3.1}$$

We can now state the main result, which holds for general settings including in particular the three examples presented in the introduction.

Theorem 1. *Let Assumption 1 be satisfied with $c_0 = 5$ and $\text{rank}(A_0) \leq r$. Then, the estimator (1.3) satisfies on the event $\lambda \geq 3\|\mathbf{M}_1 + \mathbf{M}_2\|_\infty$*

$$\|\hat{A}^\lambda - A_0\|_\infty \leq \left(\frac{5}{6} + \frac{6\sqrt{2}}{11(\alpha - 1)} \right) \frac{\lambda}{\kappa_1^2}. \quad (3.2)$$

In [13], the authors obtained an oracle inequality for the Frobenius norm with an upper bound proportional to $\text{rank}(A_0)\lambda/\kappa_1^2$ (with our notations), which trivially implies a suboptimal bound for the spectral norm since $\|\cdot\|_\infty \leq \|\cdot\|_2$. Under Assumption 1, we obtain a bound (3.2) that does not depend on $\text{rank}(A_0)$. We will see in Section 4 that this oracle inequality gives the optimal rate for the spectral norm in the matrix completion problem.

Proof. Note first that a necessary condition of extremum in the minimization problem (1.3) implies that there exists $\hat{V} \in \partial\|\hat{A}^\lambda\|_1$ such that, for all $A \in \mathbb{R}^{m_1 \times m_2}$

$$2\langle \hat{A}^\lambda, \hat{A}^\lambda - A \rangle_{L_2(\Pi)} - \left\langle \frac{2}{n} \sum_{i=1}^n Y_i X_i, \hat{A}^\lambda - A \right\rangle + \lambda \langle \hat{V}, \hat{A}^\lambda - A \rangle = 0.$$

Set $\Delta = \hat{A}^\lambda - A_0$. It follows from the previous display that, for any $U \in \mathbb{R}^{m_1 \times m_2}$ with $\|U\|_1 = 1$,

$$|\langle \Delta, U \rangle_{L_2(\Pi)}| \leq \|\mathbf{M}_1 + \mathbf{M}_2\|_\infty + \frac{\lambda}{2}.$$

Thus we get, on the event $\lambda \geq 3\|\mathbf{M}_1 + \mathbf{M}_2\|_\infty$, for any $U \in \mathbb{R}^{m_1 \times m_2}$ with $\|U\|_2 = 1$,

$$|\langle \Delta, U \rangle_{L_2(\Pi)}| \leq \frac{5}{6}\lambda. \quad (3.3)$$

Next, recall that the SVD of $\Delta = \hat{A}^\lambda - A_0$ is

$$\Delta = \sum_{j=1}^{\hat{r}} \sigma_j(\Delta) u_j^{(\Delta)} \otimes v_j^{(\Delta)}, \quad \hat{r} = \text{rank}(\Delta).$$

Take $U = u_1^{(\Delta)} \otimes v_1^{(\Delta)}$. Then, we have

$$\langle \Delta, U \rangle_{L_2(\Pi)} = \langle P_1(\Delta), U \rangle_{L_2(\Pi)} + \langle P_1^\perp(\Delta), U \rangle_{L_2(\Pi)},$$

where P_1 and P_1^\perp denote the orthogonal projections onto $M_1 = \text{l.s.} \left(u_1^{(\Delta)} \otimes v_1^{(\Delta)} \right)$ and M_1^\perp respectively. Combining the previous display with Equation (3.3) and Assumption 1 gives

$$|\langle P_1(\Delta), U \rangle_{L_2(\Pi)}| \leq \frac{5}{6}\lambda + \rho\|U\|_1\|P_1^\perp(\Delta)\|_1 \leq \frac{5}{6}\lambda + \rho\|P_1^\perp(\Delta)\|_1 \leq \frac{5}{6}\lambda + \rho\|\Delta\|_1.$$

Lemma 1 yields on the event $\lambda \geq 3\|\mathbf{M}_1 + \mathbf{M}_2\|_\infty$ that

$$\|\mathcal{P}_{A_0}^\perp(\Delta)\|_1 \leq 5\|\mathcal{P}_{A_0}(\Delta)\|_1,$$

which implies that $\Delta = \hat{A}^\lambda - A_0 \in \mathbb{C}_{A_0,5}$. Combining the last two displays, we get on the event $\lambda \geq 3\|\mathbf{M}_1 + \mathbf{M}_2\|_\infty$

$$\begin{aligned} |\langle P_1(\Delta), U \rangle_{L_2(\Pi)}| &\leq \frac{5}{6}\lambda + 6\sqrt{2\text{rank}(A_0)\rho}\|\mathcal{P}_{A_0}(\Delta)\|_2 \\ &\leq \frac{5}{6}\lambda + 6\sqrt{2\text{rank}(A_0)\rho\mu_5(A_0)}\|\Delta\|_{L_2(\Pi)}. \end{aligned}$$

Theorem 2 in [13] with $A = A_0$ gives on the event $\lambda \geq 3\|\mathbf{M}_1 + \mathbf{M}_2\|_\infty$

$$\|\Delta\|_{L_2(\Pi)} \leq \lambda\mu_5(A_0)\sqrt{\text{rank}(A_0)}.$$

Combining the last two displays, we get on the event $\lambda \geq 3\|\mathbf{M}_1 + \mathbf{M}_2\|_\infty$

$$\begin{aligned} |\langle P_1(\Delta), U \rangle_{L_2(\Pi)}| &\leq \frac{5}{6}\lambda + 6\sqrt{2\text{rank}(A_0)\rho\mu_5(A_0)^2}\lambda \\ &\leq \left(\frac{5}{6} + \frac{6\sqrt{2}}{11(\alpha-1)}\right)\lambda, \end{aligned}$$

where we have used Assumption 1 and Proposition 1 in the second line.

Next, note that

$$\langle P_1(\Delta), U \rangle_{L_2(\Pi)} = \sigma_1(\Delta)\|U\|_{L_2(\Pi)}^2 \geq \sigma_1(\Delta)\kappa_1^2\|U\|_2^2 = \sigma_1(\Delta)\kappa_1^2.$$

Finally, combining the last two displays, we get the result. \square

4. Matrix completion upper bounds with the spectral norm

In this section, we apply the general results of the previous section to the matrix completion problem with i.i.d. sub-exponential noise variables.

Assumption 2. *There exist constants $\sigma, c_1 > 0$, $\beta \geq 1$ and \tilde{c} such that*

$$\max_{i=1,\dots,n} \mathbb{E} \exp\left(\frac{|\xi_i|^\beta}{\sigma^\beta}\right) < \tilde{c}, \quad \mathbb{E}\xi_i^2 \geq \bar{c}\sigma^2, \quad \forall 1 \leq i \leq n. \quad (4.1)$$

We need the following additional condition on κ_1 and κ'_1 .

Assumption 3. *There exist constants $0 < c_1 \leq c'_1 < \infty$ such that*

$$\sqrt{\frac{c_1}{m_1 m_2}} \leq \kappa_1 \leq \kappa'_1 \leq \sqrt{\frac{c'_1}{m_1 m_2}}. \quad (4.2)$$

This assumption imposes that the probability to observe any entry is not too small or too large. It guarantees that any low-rank matrix can be estimated with optimal spectral norm rate (up to logarithmic factors). Indeed, when Assumption 3 is satisfied, we can establish that the stochastic errors $\|\mathbf{M}_1\|_\infty$ and $\|\mathbf{M}_2\|_\infty$ are small enough with probability close to 1.

Set $m = m_1 + m_2$ and $M = m_1 \vee m_2$. Denote the entries of A_0 by $a_0(i, j)$, $1 \leq i \leq m_1$, $1 \leq j \leq m_2$. We can now state our main results concerning matrix completion.

Theorem 2. *Let X_i be i.i.d. with distribution Π on \mathcal{X} defined in (1.2). Let Assumption 1 be satisfied with $c_0 = 5$ and $\text{rank}(A_0) \leq r$. Let Assumptions 2 and 3 with, in addition, $2c'_1 \leq Mc_1$. Assume that $\max_{i,j} |a_0(i, j)| \leq a$ for some constant a . For $t > 0$, consider the regularization parameter λ satisfying*

$$\lambda \geq C(\sigma \vee a) \max \left\{ \sqrt{\frac{t + \log(m)}{(m_1 \wedge m_2)n}}, \frac{(t + \log(m)) \log^{1/\beta}(m_1 \wedge m_2)}{n} \right\}, \quad (4.3)$$

where $C > 0$ is a large enough constant that can depend only on $\alpha, \beta, \bar{c}, \bar{c}, c_1, c'_1$. Then, the estimator (1.3) satisfies, with probability at least $1 - e^{-t}$

$$\|\hat{A}^\lambda - A_0\|_\infty \leq C'(\sigma \vee a) m_1 m_2 \max \left\{ \sqrt{\frac{t + \log(m)}{(m_1 \wedge m_2)n}}, \frac{(t + \log(m)) \log^{1/\beta}(m_1 \wedge m_2)}{n} \right\}, \quad (4.4)$$

where $C' > 0$ can depend only on $\alpha, \beta, \bar{c}, \bar{c}, c_1, c'_1$.

Note that the technical condition $2c'_1 \leq Mc_1$ is mild when $M \geq 2$ is large. Note also that when the noise variables are bounded, then this technical condition is no longer needed since we can apply Proposition 2 instead of Proposition 3 in Section 5 to control $\|\mathbf{M}_1\|_\infty$.

Proof. This proof consists in applying Theorem 1 with a sufficiently large λ such that the condition $\lambda \geq 3\|\mathbf{M}_1 + \mathbf{M}_2\|_\infty$ holds with probability close to 1. To this end, we need to control the stochastic errors $\|\mathbf{M}_1\|_\infty$ and $\|\mathbf{M}_2\|_\infty$; see Lemmas 2 and 3 in Section 5 below. Next a simple union bound argument gives for any λ satisfying (4.3) that (4.4) holds with probability at least $1 - 3e^{-t}$, which can then be rewritten as $1 - e^{-t}$ with a proper adjustment of the constants. \square

Note that the natural choice of t is of the order $\log(m)$. In addition, if $n > M \log^{1+2/\beta}(m)$, then we choose λ of the form

$$\lambda = C(\sigma \vee a) \sqrt{\frac{\log(m)}{(m_1 \wedge m_2)n}}, \quad (4.5)$$

where $C > 0$ is a large enough constant that can depend only on $\alpha, \beta, \bar{c}, \bar{c}, c_1, c'_1$. We immediately obtain the following corollary of Theorem 2

Corollary 1. *Let the assumptions of Theorem 2 be satisfied with λ as in (4.5) and a large enough constant $C > 0$ that can depend only on $\alpha, \beta, \tilde{c}, \bar{c}, c_1, c'_1$, $n > (m_1 \vee m_2) \log^{1+2/\beta}(m)$.*

Then, the estimator (1.3) satisfies, with probability at least $1 - 1/m$,

$$\|\hat{A}^\lambda - A_0\|_\infty \leq C'(\sigma \vee a)\sqrt{m_1 m_2} \sqrt{\frac{M \log m}{n}}, \quad (4.6)$$

where $C' > 0$ can depend only on $\alpha, \beta, \tilde{c}, \bar{c}, c_1, c'_1$.

We prove now that the above result is optimal up to logarithmic factors by establishing a minimax lower bound. We will denote by $\inf_{\hat{A}}$ the infimum over all estimators \hat{A} with values in $\mathbb{R}^{m_1 \times m_2}$. For any integer $r \leq \min(m_1, m_2)$ and any $a > 0$ we consider the class of matrices

$$\mathcal{A}(r, a) = \{A_0 \in \mathbb{R}^{m_1 \times m_2} : \text{rank}(A_0) \leq r, \max_{i,j} |a_0(i, j)| \leq a\}.$$

For any $A \in \mathbb{R}^{m_1 \times m_2}$, let \mathbb{P}_A denote the probability distribution of the observations $(X_1, Y_1, \dots, X_n, Y_n)$ with $\mathbb{E}(Y_i | X_i) = \langle A, X_i \rangle$.

Theorem 3. *Fix $a > 0$ and an integer r such that $1 \leq r \leq m_1 \wedge m_2$, $Mr \leq n$. Let the matrices X_i be i.i.d. with distribution Π on \mathcal{X} satisfying Assumption 3. Let the variables ξ_i be independent Gaussian $\mathcal{N}(0, \sigma^2)$, $\sigma^2 > 0$, for $i = 1, \dots, n$. Then there exist absolute constants $\beta \in (0, 1)$ and $c > 0$, such that*

$$\inf_{\hat{A}} \sup_{A_0 \in \mathcal{A}(r, a)} \mathbb{P}_{A_0} \left(\|\hat{A} - A_0\|_\infty > c(\sigma \wedge a)\sqrt{m_1 m_2} \sqrt{\frac{Mr}{n}} \right) \geq \beta. \quad (4.7)$$

The proof of this result can be found in Section 6 below.

Comparing Theorem 3 with Corollary 1 we see that, in the case of Gaussian errors ξ_i , the rate of convergence of \hat{A}^λ is optimal (up to a logarithmic factor) in a minimax sense on the class of matrices $\mathcal{A}(r, a)$.

5. Proofs

5.1. An intermediate result

We need the following lemma to prove Theorem 1.

Lemma 1. *The estimator (1.3) satisfies, on the event $\lambda \geq 3(\|\mathbf{M}_1 + \mathbf{M}_2\|_\infty)$*

$$\|\mathcal{P}_{A_0}^\perp(\hat{A}^\lambda - A_0)\|_1 \leq 5\|\mathcal{P}_{A_0}(\hat{A}^\lambda - A_0)\|_1.$$

Note that this result is an intermediate result in the proof of Theorem 2 in [13]. For the sake of completeness, we provide here a proof of this result.

Proof. Note that a necessary condition of extremum in the minimization problem (1.3) implies that there exists $\hat{V} \in \partial \|\hat{A}^\lambda\|_1$ such that, for all $A \in \mathbb{R}^{m_1 \times m_2}$

$$2\langle \hat{A}^\lambda, \hat{A}^\lambda - A \rangle_{L_2(\Pi)} - \left\langle \frac{2}{n} \sum_{i=1}^n Y_i X_i, \hat{A}^\lambda - A \right\rangle + \lambda \langle \hat{V}, \hat{A}^\lambda - A \rangle = 0.$$

Set $\mathbf{M} = \mathbf{M}_1 + \mathbf{M}_2$. It follows from the previous display that

$$2\|\hat{A}^\lambda - A_0\|_{L_2(\Pi)}^2 + \lambda \langle \hat{V} - V, \hat{A}^\lambda - A_0 \rangle = -\lambda \langle V, \hat{A}^\lambda - A_0 \rangle + 2\langle \mathbf{M}, \hat{A}^\lambda - A_0 \rangle,$$

for an arbitrary $V \in \partial \|A_0\|_1$. For the sake of brevity, we set $A_0 = \sum_{j=1}^r \sigma_j u_j \otimes v_j$ where $r = \text{rank}(A_0)$, $u_j = u_j^{(A_0)}$, $v_j = v_j^{(A_0)}$ and $S_j = S_j(A_0)$, $j = 1, 2$. Then, V admits the following representation

$$V = \sum_{j=1}^r u_j \otimes v_j + P_{S_1}^\perp W P_{S_2}^\perp,$$

where W is an arbitrary matrix with $\|W\|_\infty \leq 1$. By monotonicity of the sub-differential of convex functions, $\langle \hat{V} - V, \hat{A}^\lambda - A_0 \rangle \geq 0$. Therefore, we get

$$\lambda \langle P_{S_1}^\perp W P_{S_2}^\perp, \hat{A}^\lambda - A_0 \rangle \leq -\lambda \left\langle \sum_{j=1}^r u_j \otimes v_j, \hat{A}^\lambda - A_0 \right\rangle + 2\langle \mathbf{M}, \hat{A}^\lambda - A_0 \rangle.$$

Set $\Delta = \hat{A}^\lambda - A_0$. The trace duality guarantees the existence of a matrix W with $\|W\|_\infty$ such that

$$\langle P_{S_1}^\perp W P_{S_2}^\perp, \Delta \rangle = \langle W, P_{S_1}^\perp \Delta P_{S_2}^\perp \rangle = \|P_{S_1}^\perp \Delta P_{S_2}^\perp\|_1.$$

The trace duality again implies that

$$\left| \left\langle \sum_{j=1}^r u_j \otimes v_j, \Delta \right\rangle \right| \leq \|P_{S_1} \Delta P_{S_2}\|_1.$$

Combining the last three displays, we get, on the event $\lambda \geq 3(\|\mathbf{M}_1 + \mathbf{M}_2\|_\infty)$

$$\begin{aligned} \|P_{S_1}^\perp \Delta P_{S_2}^\perp\|_1 &\leq \|P_{S_1} \Delta P_{S_2}\|_1 + \frac{2}{3} \|\Delta\|_1 \\ &\leq \frac{5}{3} \|P_{S_1} \Delta P_{S_2}\|_1 + \frac{2}{3} \|P_{S_1}^\perp \Delta P_{S_2}^\perp\|_1. \end{aligned}$$

Thus we get the result. \square

5.2. Control of the stochastic errors

The following proposition is an immediate consequence of the matrix version of Bernstein's inequality (Corollary 9.1 in [22]). For the sake of brevity, we write $\|\cdot\|_\infty = \|\cdot\|$.

Proposition 2. Let Z_1, \dots, Z_n be independent random matrices with dimensions $m_1 \times m_2$ that satisfy $\mathbb{E}(Z_i) = 0$ and $\|Z_i\| \leq U$ almost surely for some constant U and all $i = 1, \dots, n$. Define

$$\sigma_Z = \max \left\{ \left\| \frac{1}{n} \sum_{i=1}^n \mathbb{E}(Z_i Z_i^\top) \right\|^{1/2}, \left\| \frac{1}{n} \sum_{i=1}^n \mathbb{E}(Z_i^\top Z_i) \right\|^{1/2} \right\}.$$

Then, for all $t > 0$, with probability at least $1 - e^{-t}$ we have

$$\left\| \frac{Z_1 + \dots + Z_n}{n} \right\| \leq 2 \max \left\{ \sigma_Z \sqrt{\frac{t + \log(m)}{n}}, U \frac{t + \log(m)}{n} \right\},$$

where $m = m_1 + m_2$.

Furthermore, it is possible to replace the L_∞ -bound U on $\|Z\|$ in the above inequality by bounds on the weaker ψ_β -norms of $\|Z\|$ defined by

$$U_Z^{(\beta)} = \inf \left\{ u > 0 : \mathbb{E} \exp(\|Z\|^\beta / u^\beta) \leq 2 \right\}, \quad \beta \geq 1.$$

Proposition 3. Let Z, Z_1, \dots, Z_n be i.i.d. random matrices with dimensions $m_1 \times m_2$ that satisfy $\mathbb{E}(Z) = 0$. Suppose that $U_Z^{(\beta)} < \infty$ for some $\beta \geq 1$. Then there exists a constant $C > 0$ such that, for all $t > 0$, with probability at least $1 - e^{-t}$

$$\left\| \frac{Z_1 + \dots + Z_n}{n} \right\| \leq C \max \left\{ \sigma_Z \sqrt{\frac{t + \log(m)}{n}}, U_Z^{(\beta)} \left(\log \frac{U_Z^{(\beta)}}{\sigma_Z} \right)^{1/\beta} \frac{t + \log(m)}{n} \right\},$$

where $m = m_1 + m_2$.

This is an easy consequence of Proposition 2 in [11], which provides an analogous result for Hermitian matrices Z . Its extension to rectangular matrices stated in Proposition 3 is straightforward via the self-adjoint dilation; see, for example, the proof of Corollary 9.1 in [22].

Lemma 2. Let the noise variables ξ_1, \dots, ξ_n be i.i.d. and satisfy Assumption 2. Let X, X_1, \dots, X_n be i.i.d. with distribution Π on \mathcal{X} satisfying Assumption 3. Then there exists an absolute constant $C > 0$ that can depend only on $\beta, \tilde{c}, \bar{c}, c_1, c'_1$ and such that, for all $t > 0$, with probability at least $1 - 2e^{-t}$ we have

$$\|\mathbf{M}_1\| \leq C\sigma \max \left\{ \sqrt{\frac{t + \log(m)}{(m_1 \wedge m_2)n}}, \frac{(t + \log(m)) \log^{1/\beta}(m_1 \wedge m_2)}{n} \right\}. \quad (5.1)$$

The proof of this lemma is essentially the same as that of Lemma 2 in [13] up to some additional technicalities due to the fact Π is no longer assumed to be the uniform distribution on \mathcal{X} . We set $\pi(i, j) = \Pi(e_i(m_1) e_j^\top(m_2))$ for any $1 \leq i \leq m_1, 1 \leq j \leq m_2$.

Proof. Clearly, we have $\|X\| = 1$. Furthermore, under Assumption 3, we have

$$\max \{ \|\mathbb{E}(X)\|, \|\mathbb{E}(X)^\top\| \} \leq \sqrt{\frac{c'_1}{m_1 m_2}}, \quad \frac{c_1}{m_1 \wedge m_2} \leq \sigma_X^2 \leq \frac{c'_1}{m_1 \wedge m_2}. \quad (5.2)$$

Indeed, Assumption 3 implies that

$$0 < \frac{c_1}{m_1 m_2} \leq \pi(i, j) \leq \frac{c'_1}{m_1 m_2}, \quad \forall i, j. \quad (5.3)$$

Next, we have

$$\|\mathbb{E}(X)\| = \max_{x \in \mathbb{R}^{m_2}: |x|_2=1} \sqrt{\sum_i \left(\sum_j \pi(i, j) x_j \right)^2}.$$

Note that the maximum is clearly achieved at point x satisfying $x_j \geq 0$ for any $1 \leq j \leq m_2$ since $\pi(i, j) > 0$ for any i, j in view of the two above displays. Thus, we get

$$\begin{aligned} \|\mathbb{E}(X)\| &\leq \sqrt{\frac{c'_1}{m_1 m_2}} \max_{x \in \mathbb{R}^{m_2}: |x|_2=1} \sqrt{\sum_i \left(\sum_j \sqrt{\pi(i, j)} x_j \right)^2} \\ &\leq \sqrt{\frac{c'_1}{m_1 m_2}} \max_{x \in \mathbb{R}^{m_2}: |x|_2=1} \sqrt{\sum_i \left(\sum_j \pi(i, j) \right) \left(\sum_j x_j^2 \right)} \\ &\leq \sqrt{\frac{c'_1}{m_1 m_2}} \sqrt{\sum_i \left(\sum_j \pi(i, j) \right)} \leq \sqrt{\frac{c'_1}{m_1 m_2}}, \end{aligned}$$

where we have used successively Cauchy-Schwarz's inequality, $|x|_2 = 1$ and $\sum_{i,j} \pi(i, j) = 1$. Similarly, We obtain the same bound for $\|\mathbb{E}(X)^\top\|$.

We have

$$\sigma_X^2 = \max \left\{ \max_{1 \leq i \leq m_1} \left(\sum_{j=1}^{m_2} \pi(i, j) \right), \max_{1 \leq j \leq m_2} \left(\sum_{i=1}^{m_1} \pi(i, j) \right) \right\}.$$

Combining the above display with (5.3) yields the second part of (5.2).

Next, observe that for $\tilde{X} = X - \mathbb{E}(X)$, we have in view of (5.2) and the technical condition $2c'_1 \leq M c_1$ that

$$\frac{c_1}{2m_1 \wedge m_2} \leq \sigma_{\tilde{X}}^2 \leq \frac{2c'_1}{m_1 \wedge m_2}. \quad (5.4)$$

Indeed, this follows from the easy fact

$$\|\mathbb{E}(XX^\top)\| - \|\mathbb{E}(X)\| \|\mathbb{E}(X)^\top\| \leq \|\mathbb{E}(\tilde{X}\tilde{X}^\top)\| \leq \|\mathbb{E}(XX^\top)\| + \|\mathbb{E}(X)\| \|\mathbb{E}(X)^\top\|,$$

combined with (5.2) and Assumption 3. We proceed similarly for $\|\mathbb{E}\tilde{X}^\top\tilde{X}\|$.

Now,

$$\begin{aligned} \|\mathbf{M}_1\| &\leq \left\| \frac{1}{n} \sum_{i=1}^n \xi_i (X_i - \mathbb{E}X_i) \right\| + \left\| \frac{1}{n} \sum_{i=1}^n \xi_i \mathbb{E}(X_i) \right\| \\ &\leq \left\| \frac{1}{n} \sum_{i=1}^n \xi_i (X_i - \mathbb{E}X) \right\| + \sqrt{\frac{c'_1}{m_1 m_2}} \left| \frac{1}{n} \sum_{i=1}^n \xi_i \right|. \end{aligned} \quad (5.5)$$

Set $Z_i = \xi_i (X_i - \mathbb{E}X)$. These are i.i.d. random matrices having the same distribution as a random matrix Z . Since $\|X\| = 1$ we have that $\|Z_i\| \leq 2|\xi_i|$, and thus Assumption 2 implies that $U_Z^{(\beta)} \leq c\sigma$ for some constant $c > 0$. Furthermore, in view of (5.4), we have $\sigma_Z \leq c_2\sigma/(m_1 \wedge m_2)^{1/2}$ for some constant $c_2 > 0$ depending only on c'_1 and $\sigma_Z \geq c_3\sigma/(m_1 \wedge m_2)^{1/2}$ for some constant $c_3 > 0$ depending only on c_1, \bar{c} . Using these remarks we can deduce from Proposition 3 that there exists an absolute constant $\tilde{C} > 0$ such that for any $t > 0$ with probability at least $1 - e^{-t}$ we have

$$\begin{aligned} &\left\| \frac{1}{n} \sum_{i=1}^n \xi_i (X_i - \mathbb{E}X) \right\| \\ &\leq \tilde{C} \max \left\{ \sigma_Z \sqrt{\frac{t + \log(m)}{n}}, \quad U_Z^{(\beta)} \left(\log \frac{U_Z^{(\beta)}}{\sigma_Z} \right)^{1/\beta} \frac{t + \log(m)}{n} \right\} \\ &\leq C\sigma \max \left\{ \sqrt{\frac{t + \log(m)}{(m_1 \wedge m_2)n}}, \quad \frac{(t + \log(m)) \log^{1/\beta}(m_1 \wedge m_2)}{n} \right\}. \end{aligned}$$

Finally, in view of Assumption (2) and Bernstein's inequality for sub-exponential noise, we have for any $t > 0$, with probability at least $1 - e^{-t}$,

$$\left| \frac{1}{n} \sum_{i=1}^n \xi_i \right| \leq C\sigma \max \left\{ \sqrt{\frac{t}{n}}, \frac{t}{n} \right\},$$

where $C > 0$ depends only on \tilde{c} . We complete the proof by using the union bound. \square

We now treat $\|\mathbf{M}_2\|$.

Lemma 3. *Let X, X_1, \dots, X_n be i.i.d. random variables with distribution Π on \mathcal{X} satisfying Assumption 3. Assume, in addition, that $\max_{i,j} |a_0(i, j)| \leq a$ for some $a > 0$. Then, for all $t > 0$, with probability at least $1 - e^{-t}$ we have*

$$\|\mathbf{M}_2\| \leq 2c'_1 a \max \left\{ \sqrt{\frac{t + \log(m)}{(m_1 \wedge m_2)n}}, \quad \frac{2(t + \log(m))}{n} \right\}. \quad (5.6)$$

Proof. We apply Proposition 2 for the random variables $Z_i = \text{tr}(A_0^\top X_i)X_i - \mathbb{E}(\text{tr}(A_0^\top X)X)$. Using (5.2) we get $\|Z_i\| \leq 2 \max_{i,j} |a_0(i,j)|$ and

$$\sigma_Z^2 \leq \max \{ \|\mathbb{E}(\langle A_0, X \rangle^2 X X^\top)\|, \|\mathbb{E}(\langle A_0, X \rangle^2 X^\top X)\| \} \leq a^2 \frac{c'_1}{m_1 \wedge m_2}.$$

Thus, (5.6) follows from Proposition 2. \square

5.3. Proof of Theorem 3

Proof. We assume w.l.o.g. that $M = m_1 \vee m_2 = m_1 \geq m_2$. The idea is to adapt to our context Theorem 5 in [13]. Note that Theorem 5 is established under a restricted isometry condition in expectation (See Assumption 2 in [13]). A quick investigation of the proof shows that the conclusion of this theorem is still valid for X_1, \dots, X_n i.i.d. with distribution Π satisfying Assumption 3. Indeed, we then have for any $A \in \mathbb{R}^{m_1 \times m_2}$

$$\frac{c_1}{m_1 m_2} \|A\|_2^2 \leq \|A\|_{L_2(\Pi)}^2 \leq \frac{c'_1}{m_1 m_2} \|A\|_2^2, \quad (5.7)$$

Recall that [13] established in the proof of Theorem 5 the existence of a subset $\mathcal{A}^0 \subset \mathcal{A}(r, a)$ with cardinality $\text{Card}(\mathcal{A}^0) \geq 2^{r m_1/8} + 1$ containing the zero $m_1 \times m_2$ matrix $\mathbf{0}$ and such that, for any two distinct elements A_1 and A_2 of \mathcal{A}^0 ,

$$\frac{\gamma^2}{16} (\sigma \wedge a)^2 \frac{m_1^2 m_2 r}{n} \leq \|A_1 - A_2\|_2^2 \leq \gamma^2 (\sigma \wedge a)^2 \frac{m_1^2 m_2 r}{n}. \quad (5.8)$$

Next, using (5.7) instead of Assumption 2 in [13], Equations (4.3) and (4.4) in [13] are replaced respectively by

$$\|A_1 - A_2\|_{L_2(\Pi)}^2 \geq c_1 \frac{\gamma^2}{16} (\sigma \wedge a)^2 \frac{m_1 r}{n}, \quad (5.9)$$

and

$$K(\mathbb{P}_0, \mathbb{P}_A) = \frac{n}{2\sigma^2} \|A\|_{L_2(\Pi)}^2 \leq c'_1 \frac{\gamma^2}{2} m_1 r, \quad (5.10)$$

where $K(\mathbb{P}_0, \mathbb{P}_A)$ is the Kullback-Leibler distance between \mathbb{P}_0 and \mathbb{P}_A and $\gamma > 0$ is some numerical quantity introduced in the construction of the set \mathcal{A}^0 in [13].

For any two distinct matrices A_1, A_2 of \mathcal{A}^0 , we have

$$\|A_1 - A_2\|_\infty \geq \sqrt{\frac{c_1}{c'_1}} \sqrt{\frac{\gamma}{16}} (\sigma \wedge a) \sqrt{\frac{m_1^2 m_2}{n}}. \quad (5.11)$$

Indeed, if (5.11) does not hold, we get

$$\|A_1 - A_2\|_{L_2(\Pi)}^2 \leq \frac{c_1}{m_1 m_2} \text{rank}(A_1 - A_2) \|A_1 - A_2\|_\infty^2 < c_1 \frac{\gamma}{16} (\sigma \wedge a)^2 \frac{m_1 r}{n},$$

since $\text{rank}(A_1 - A_2) \leq r$ by construction of \mathcal{A}^0 in [13]. This contradicts (5.9).

We now take $\gamma > 0$ sufficiently small depending only on c'_1, c_1, α with $\alpha > 0$ so that

$$\frac{1}{\text{Card}(\mathcal{A}^0) - 1} \sum_{A \in \mathcal{A}^0} K(\mathbb{P}_0, \mathbb{P}_A) \leq \alpha \log(\text{Card}(\mathcal{A}^0) - 1). \quad (5.12)$$

Combining (5.11) with (5.12) and Theorem 2.5 in [23] gives the result. \square

References

- [1] ARGYRIOU, A., EVGENIOU, T. AND PONTIL, M. (2008) Convex multi-task feature learning. *Machine Learning*, **73**, 243–272.
- [2] BUNEA, F. (2008) Consistent selection via the Lasso for high dimensional approximating regression models. In *Pushing the limits of contemporary statistics: contributions in honor of Jayanta K. Ghosh*, volume 3 of *Inst. Math. Stat. Collect.*, pages 122–137. Inst. Math. Statist., Beachwood, OH.
- [3] BUNEA, F. AND TSYBAKOV, A.B. AND WEGKAMP, M.H. (2007) Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics*, 1:169–194.
- [4] CANDÈS, E. J. AND PLAN, Y. (2009) Matrix completion with noise. *Proceedings of IEEE*.
- [5] CANDÈS, E. J. AND PLAN, Y. (2010) Tight oracle bounds for low-rank matrix recovery from a minimal number of noisy random measurements. [arXiv:1001.0339](#). January, 2010.
- [6] CANDÈS, E. J. AND RECHT, B. (2009) Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6), 717–772.
- [7] CANDÈS, E. AND TAO, T. (2009) The power of convex relaxation: Near-optimal matrix completion. [arXiv:0903.1476](#)
- [8] DONOHO, D.L. AND ELAD, M. AND TEMLYAKOV, V.N. (2006) Stable recovery of sparse overcomplete representations in the presence of noise. *Information Theory, IEEE Transactions on*, 52(1):6–18.
- [9] GROSS, D. (2009) Recovering low-rank matrices from few coefficients in any basis. [arXiv:0910.1879](#).
- [10] KESHAVAN, R.H., MONTANARI, A. AND OH, S. (2009) Matrix completion from noisy entries. [arXiv:0906.2027](#)
- [11] KOLTCHINSKII, V. (2010) von Neumann entropy penalization and low rank matrix approximation. [arXiv:1009.2439](#)
- [12] KOLTCHINSKII, V. (2011) Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems. *École d'Été de Probabilités de Saint-Flour XXXVIII-2008*. Series: Lecture Notes in Mathematics, Vol. 2033
- [13] KOLTCHINSKII, V., LOUNICI, K. AND TSYBAKOV, A.B. (2010) Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.*, to appear.
- [14] LOUNICI, K. (2008) Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electronic Journal of Statistics*, 2:90–102.
- [15] LOUNICI, K., PONTIL, M., TSYBAKOV, A.B. AND VAN DE GEER, S.A. (2010) Oracle inequalities and optimal inference under group sparsity *Ann. Statist.*, to appear.
- [16] NEGAHBAN, S. AND WAINWRIGHT, M.J. (2009) Estimation of (near) low rank matrices with noise and high-dimensional scaling. [arXiv:0912.5100](#), December 2009.
- [17] NEGAHBAN, S. AND WAINWRIGHT, M.J. (2010) Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. [arXiv:1009.2118](#)
- [18] NEGAHBAN, S., RAVIKUMAR, P., WAINWRIGHT, M.J., AND YU, B. (2010) A unified framework for high-dimensional analysis of M -estimators with decomposable regularizers. [arXiv:1010.2731](#)
- [19] RECHT, B. (2009) A simpler approach to matrix completion. [arXiv:0910.0651](#)

- [20] RECHT, B., FAZEL, M. AND PARRILO, P.A. (2007) Guaranteed Minimum-Rank Solutions of Linear Matrix Equations via Nuclear Norm Minimization. [arXiv:0706.4138](#)
- [21] ROHDE, A. AND TSYBAKOV, A. (2009) Estimation of high-dimensional low rank matrices. [arXiv:0912.5338](#) December 2009.
- [22] TROPP, J. A. (2010) User-friendly tail bounds for sums of random matrices. [arXiv:1004.4389](#), April 2010.
- [23] TSYBAKOV, A. (2009) *Introduction to Nonparametric Estimation*. Springer.
- [24] WATSON, G. A. (1992) Characterization of the subdifferential of some matrix norms. *Linear Algebra Appl.*, 170, 33-45.